Submodular Feature Selection for High-Dimensional Acoustic Score Space

Yuzong Liu[1], Kai Wei[1], Katrin Kirchhoff[1,2], Yisong Song[2], Jeff Bilmes[1,2]

[1] Department of Electrical Engineering [2] Department of Computer Science and Engineering
University of Washington, Seattle, USA

UNIVERSITY of WASHINGTON

1. Background

- Generative acoustic score spaces, such as the Fisher score space, are widely used in speech processing, including acoustic event classification, acoustic-phonetic classification, segmental minimum Bayes risk decoding, and speaker verification. The drawback of these score space is their high dimensionality.
- This work presents a general-purpose feature selection method based on submodular function maximization. The problem can be constant-factor approximated with a simple scalable accelerated greedy algorithm.

2. Fisher Score Spaces

- Fisher score vectors: contain derivatives of data log-likelihood w.r.t. the parameters of a generative model

$$U_X^\theta = \nabla_\theta log P(X|\theta)$$: acoustic data, θ: parameters

- When multiple models are involved, Fisher score vectors for each model are stacked to form complete score space:

$$U'_X = ((U_X^{\theta_1})^\intercal, (U_X^{\theta_2})^\intercal, ..., (U_X^{\theta_n})^\intercal)^\intercal$$

- Often used to compute Fisher kernel similarity measure:

$$K_{i,j} = U_i' F^{-1} U_j'$$ F : Fisher information matrix

- Problem with Fisher score spaces:
  * extremely high-dimensional (e.g. 48 HMMs with 16-component Gaussian mixtures each => >180k dimensions)
  * computationally inefficient
  * many dimensions may be noisy/uninformative
- Previous Approaches to dimensionality reduction in Fisher kernels:
  1) Selectively use some dimensions (i.e. means, diagonal covariance matrices)
  2) Binary compression
  3) Feature selection using mutual information (modular rank-and-select approach)

3. Submodular Feature Selection

a) Submodular Functions

- Class of discrete functions that have a *diminishing returns* property
- Given a finite set V, a function $f: 2^V \to \mathbb{R}$ is ***submodular*** if for any $R \subseteq S \subseteq V$ and $k \in V \setminus S$:

$$f(S+k) - f(S) \le f(R+k) - f(R)$$

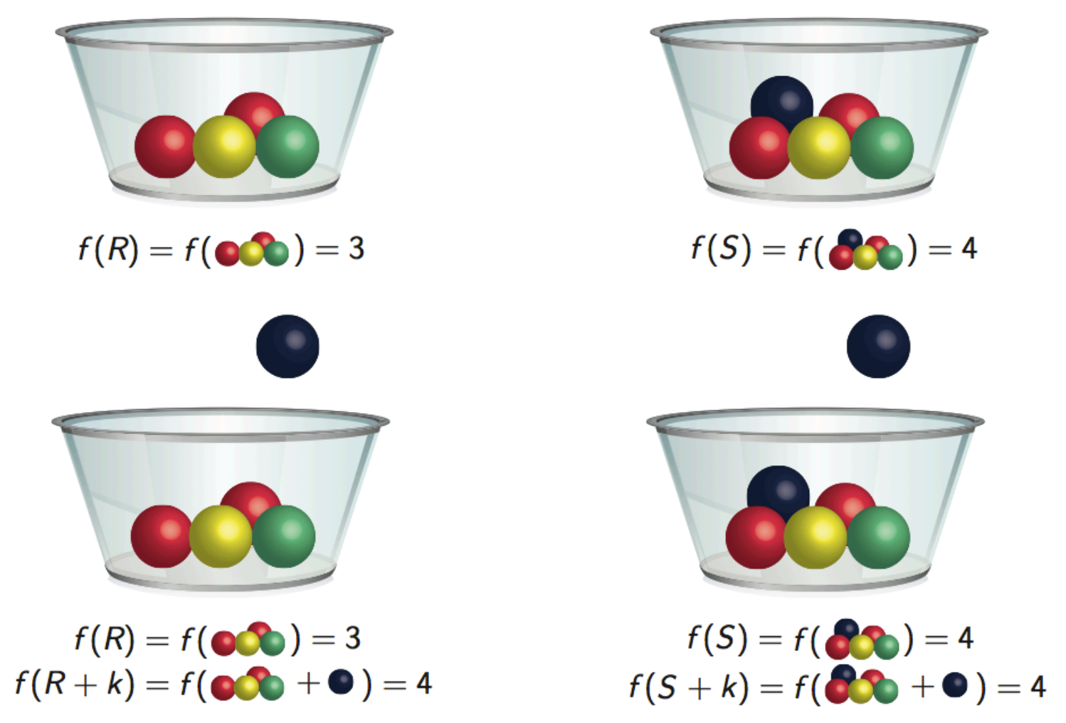

Figure 1: An example of the diminishing returns property in submodular functions.

b) Submodular Functions for Feature Selection

Problem Formulation

- A set of features $V = \{f_1, f_2, \cdots, f_n\}$;
- A submodular function $f: 2^V \to \mathbb{R}$ measures the quality of feature subset S;
- K is the total number of features to be selected.

Optimization problem: $S^* = \underset{S \subseteq V}{\arg\max} f(S)$ subject to $|S| \le K$

Submodular Function Instantiations

* Facility location function:

$$\mathcal{L}_{fl}(S) = \sum_{i \in V} \max_{j \in S} w_{ij}$$

indicates how well each feature $i \in V$ is represented by the selected subset S and w_{ij} is the mutual information between feature *i* and feature *j*.

* Saturated coverage function:

$$\mathcal{L}_{sc}(S) = \sum_{i \in V} \min\{C_i(S), \beta C_i(V)\} \quad \text{with} \quad C_i(S) = \sum_{j \in S} w_{ij}$$

- $C_i(S)$ measures the degree to which feature $i \in V$ is covered by S.
- β is a hyperparameter that determines a saturation threshold, such that the features are not over-represented by the selected subset S.
- If saturated, increasing $C_i(S)$ will not increase the value of the function. Thus, the function is forced to pick features that are not yet saturated.

Accelerated Greedy Algorithm

* Greedy algorithm can be used to solve the optimization with near-optimal solution.
* Scalable to high-dimensional feature spaces with an accelerated greedy algorithm

4. Tasks, Data Sets and Baseline Systems

a) Evaluation Tasks:

- **Task 1: Data subset selection for phone recognizer training: find a subset of the original training dataset without significant reduction in performance**

- Data subset selection: also done using submodular selection; requires a graph with similarity weights.
- Similarity is computed by Fisher kernel; goal is to reduce dimensionality of Fisher scores to improve graph and thereby the data selection results.

- **Task 2: Graph-based semi-supervised learning (SSL) for phone segment classification**

- Graph-based SSL also requires a similarity-weighted graph; goal is to use dimensionality reduction to improve Fisher kernel and thereby classification results.
- Both tasks involve a high-dimensional acoustic feature space (dimensionality > 180k).
- Experimental evaluation is conducted on the TIMIT dataset.

b) Baseline Systems

- Task 1: Data Subset Selection
  - dimensionality of Fisher score vectors: 186,577
  - baseline feature selection method: use top N features with highest mutual information between feature and phonetic class
- Task 2: Graph-based SSL for phone segment classification
  - dimensionality of Fisher score vectors: 182,017
  - baseline segment classifier: HMM, accuracy = 68.02% (TIMIT core test set)
  - graph-based learner: measure propagation[1]

5. Results

Two-stage feature selection strategy

Let $V = \{f_1, f_2, \cdots, f_n\}$ be the set of all features in the Fisher score vectors.

Stage 1: prune away features whose mutual information is less than $\tau = 0.01$;

Stage 2: apply submodular feature selection (in Task 1, saturated coverage function; in Task 2, facility location function).

a) Task 1: Data Subset Selection

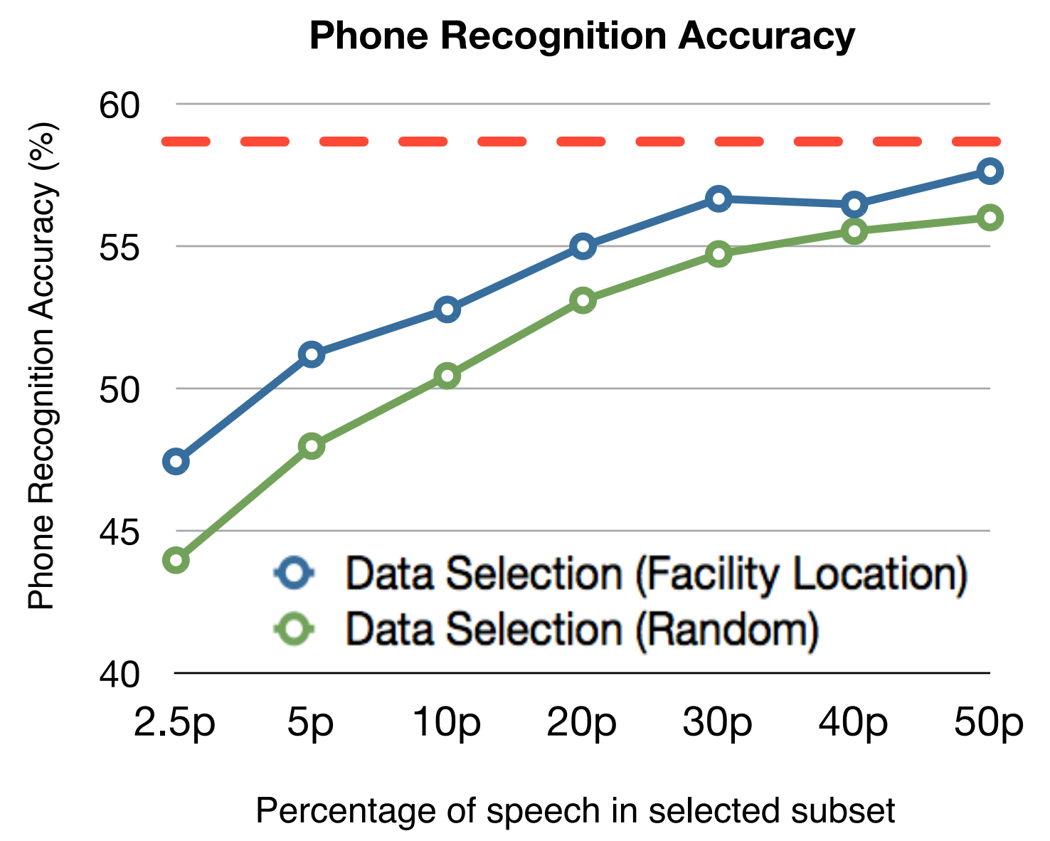

Figure 2 Phone accuracy for random subset selection and submodular subset selection using the entire Fisher score space. All (100%) of data (red), facility location (blue), and average (of 100) random selection (green).

	2.5p	5p	10p	20p	30p
1k-s	**1.99%**	**2.17%**	**2.48%**	**1.48%**	**1.93%**
1k-m	-0.20%	-0.49%	1.48%	0.52%	1.52%
2k-s	**0.11%**	**1.60%**	**2.33%**	**1.46%**	**1.97%**
2k-m	0.04%	-0.59%	0.45%	1.24%	1.45%
5k-s	0.84%	1.15%	**2.64%**	**1.12%**	**1.37%**
5k-m	**1.06%**	**1.46%**	-0.41%	-0.30%	-0.18%
10k-s	**5.21%**	**3.05%**	**3.08%**	**1.16%**	**2.84%**
10k-m	2.79%	1.29%	1.03%	0.97%	0.92%
20k-s	**6.45%**	**3.36%**	**4.34%**	**2.58%**	**1.46%**
20k-m	3.77%	2.34%	3.61%	0.79%	1.16%
50k-s	**4.79%**	**5.27%**	**3.92%**	**2.79%**	**2.71%**
50k-m	2.55%	3.85%	2.48%	1.38%	2.11%
all	5.21%	4.96%	4.04%	2.92%	2.38%

Table 1 Relative improvement in phone accuracy for different data subset sizes, different number of features, and modular (m) vs. submodular (s) feature selection methods. In 28 out of 30 settings, submodular feature selection outperforms modular selection and outperforms the full feature set in 5 settings. Submodular function = saturated coverage function.

b) Task 2: Segment Classification

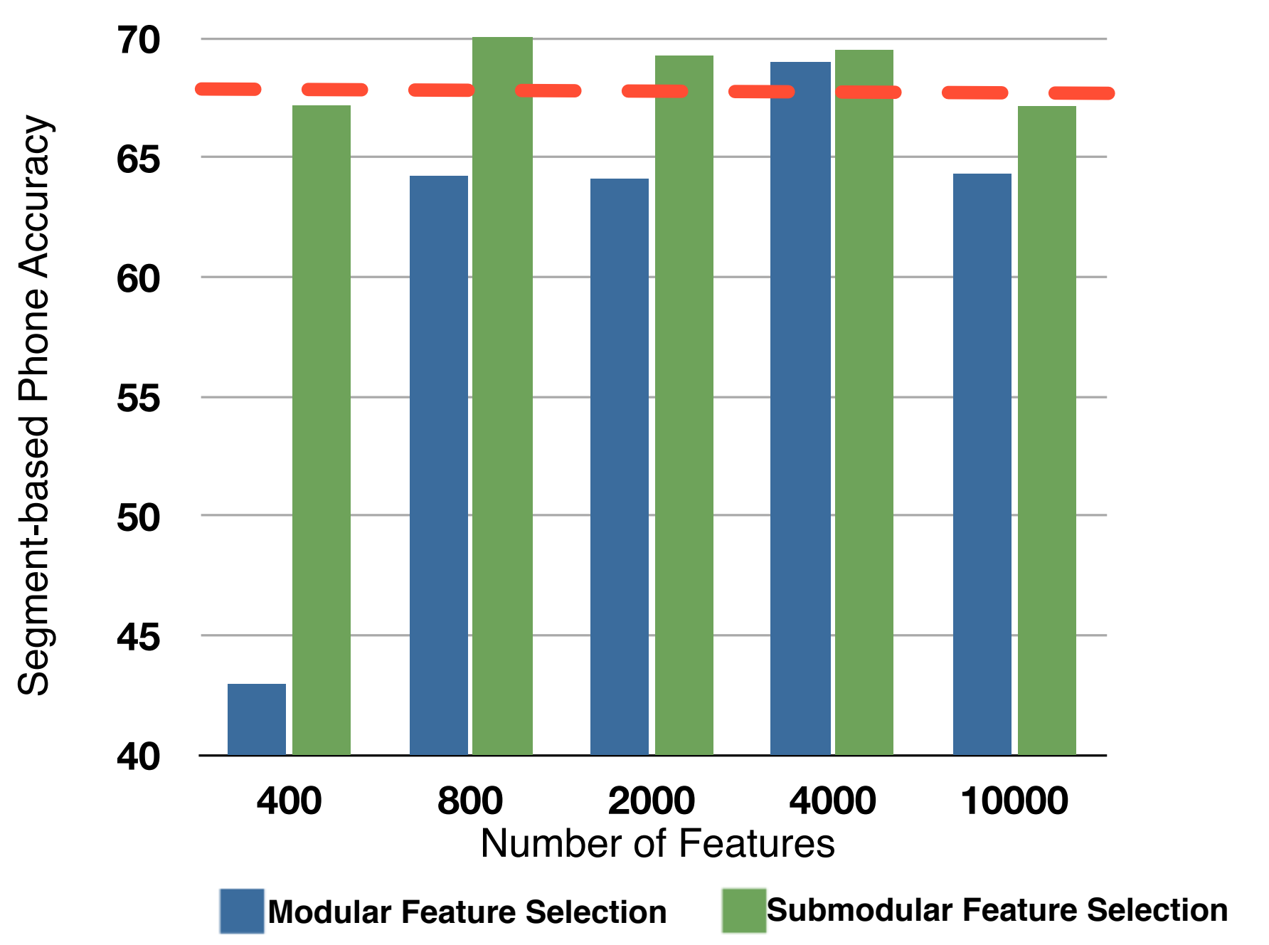

Figure 3 Phone segment classification accuracy with modular (blue) and submodular (green) feature selection. Baseline model (in red dotted line, monophone HMMs without graph-based SSL) accuracy: 68.02%. Submodular function = facility location function.

References

[1] A. Subramanya and J.A. Bilmes, *Entropic Graph Regularization in Non-Parametric Semi-Supervised Classification*, In NIPS' 09

Acknowledgment

This material is based on research sponsored by Intelligence Advanced Research Projects Activity (IARPA) under agreement number FA8650-12-2-7263.